

# Update on 2020 Census Detailed Demographic and Housing Characteristics File (Detailed DHC)

**Rachel Marks, Alexandra Krause, Nicholas Jones,  
Population Division**

**Sam Haney and William Sexton  
Tumult Labs, Office of the Deputy Director**

**Scott Holan  
Research and Methodology Directorate**

Spring 2022 Census Scientific Advisory Committee Meeting  
March 17, 2022

# Proposed 2020 Census Detailed DHC

- Separated the development of the Detailed DHC into three parts based on the types of differentially private algorithms
  - SafeTab-P: Tables T1 and T2
  - SafeTab-H: Tables T3 and T4
  - PHSafe: Complex person-household join (Join) tables
- The proposed tables were developed based on stakeholder feedback, combined with review of table access rates, and subject matter expert knowledge
- We are prioritizing the development and production of SafeTab-P (Tables T1 and T2) based on public feedback

# **Update on T1. Total Population and T2. Sex by Selected Age Categories**

# Proposed 2020 Census Detailed DHC – T1 and T2

Detailed race, ethnicity, and American Indian and Alaska Native tribes and villages

- Subjects repeated by detailed race, ethnicity, and American Indian and Alaska Native (AIAN) tribes and villages include:
  - Total population (Table T1)
  - Sex by age for selected age categories (Table T2)
- Proposed geographic levels:
  - Nation, state, county, AIANNH areas

# Timeline for Developing T1 and T2

## **Spring 2020**

- DEMO T1 and T2 team introduced to Detailed DHC Differential Privacy and Tumult Labs researchers
- Developed analytical tools for evaluating accuracy confidentiality balance

## **Summer – Fall 2020**

- DEMO researchers evaluated 2010 Census data and how best to achieve balance of privacy and accuracy

## **Fall 2020**

- DEMO presented proposal and recommendation to Census DSEP and were given approval to continue researching

## **Winter – Summer 2021**

- DEMO put T1 and T2 work on hold to review 2020 Census data files for apportionment and redistricting

## **Fall 2021 – Winter 2022**

- DEMO researchers re-engaging on T1 and T2 with R&M and Tumult after completing 2010 Census PL release
- Ongoing engagements with data users and public on Detailed DHC proposal and use cases
- Reviewing feedback from public on use cases for Detailed DHC data

# DEMO Research Goals for T1 and T2

**Develop parameter settings for algorithm to produce differentially private 2020 Census data for detailed race and ethnicity groups, and their demographic characteristics**

- Explore how to establish privacy and accuracy levels to produce statistics for our nation's myriad detailed racial and ethnic population groups
- Allow for equity across all major race and ethnicity categories
- Provide same amount, or more data, than provided from 2010 Census

# Public Feedback on T1 and T2

**Critical need for T1 and T2 statistics**

**Use cases for lower levels of geography (e.g., census tracts)**

**Ongoing engagement with stakeholders and partners about Detailed DHC development and plans for iterative feedback and discussions**

# Data Product Comparisons: 2010 vs 2020

Topics	2010	2020
Number of Population Tables	57	4
Number of Geographies	104	4
Threshold for National Sex by Age Characteristics	100	500
Threshold for National Total Count for <ul style="list-style-type: none"> <li>• AIAN Tribes and Villages</li> <li>• Detailed Race and Ethnicity Groups</li> </ul>	0 100	10 10
Number of Race and Ethnicity Groups (not inclusive of AIAN tribes and villages)	67	334
Disclosure Avoidance Technique	Swapping	Differential Privacy

Note: Estimates based on 2010 Census SF-2 data, 2016 ACS 5-Year data, and 2020 Census proposal for the Detailed DHC.



# Next Steps for Developing T1 and T2

**In March 2022, Census Bureau will share timeline for Detailed DHC development with the public**

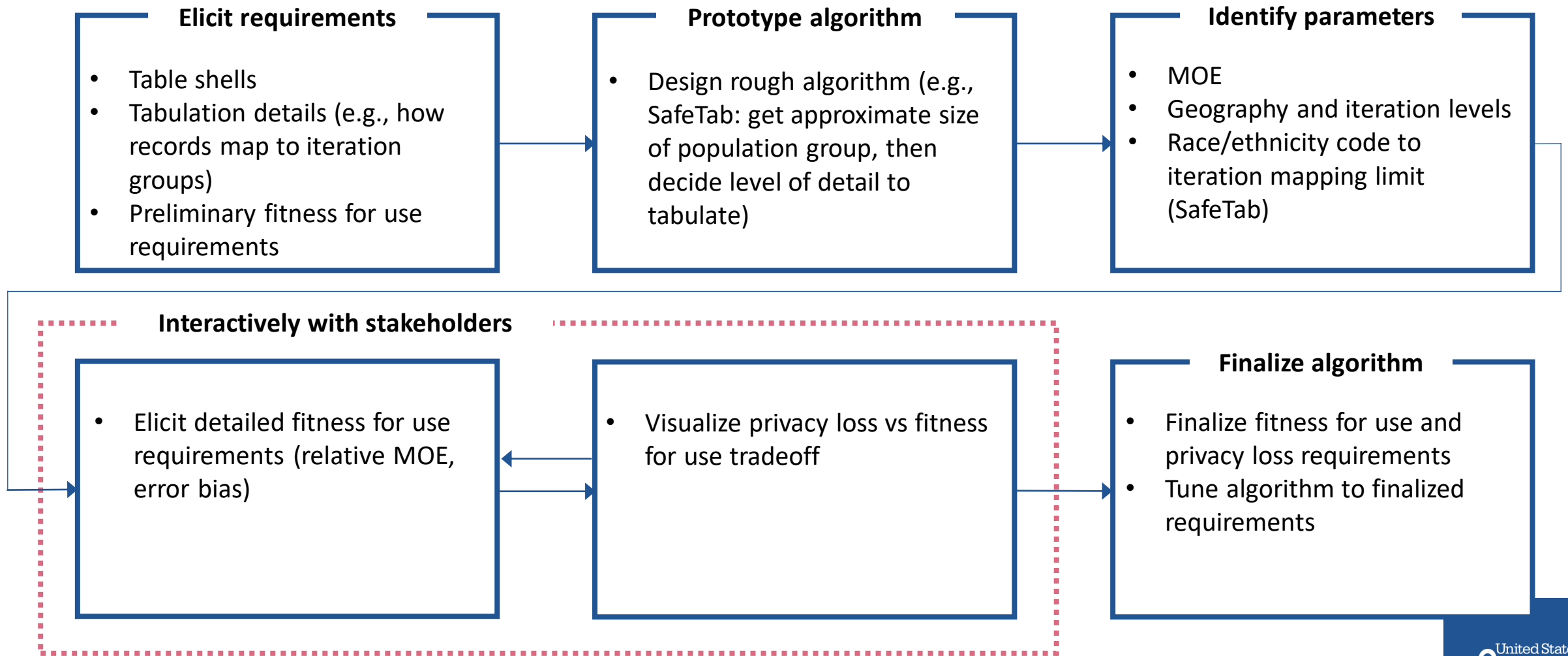
- Separating Detailed DHC into multiple pieces to prioritize T1 and T2
- Iterative process for research, public feedback, and updates to data product proposal
- Proof of concept “demonstration product”
- Establish target for data product release date

**Before final parameter settings and privacy loss budget decisions are made...**

- Additional discussions with tribal leaders through formal Tribal Consultations, as well as engagement with other stakeholders and partners
- Address feedback and recommendations from Census Bureau Advisory Committees
- Assess implications that post-processing will have on production of detailed race and ethnicity data products

# SafeTab-P: A differential privacy algorithm for T1 and T2

# Methodology



# 2020 DAS Methodology for Tables T1 and T2 - Data on Detailed Race and Ethnicity Groups and AIAN Populations

## Step 1:

Input Microdata (CEF)  
& Geographic  
Reference File (GRF-C)

## Step 2:

Convert Input  
Microdata to  
Required Format

## Step 3:

Differentially Private  
Algorithms & Noisy  
Measurements

## Step 4:

Statistical Post-  
Processing

## Step 5:

Publication of Tabular  
Summaries without  
Microdata

# Differentially Private Algorithms for Tables T1 and T2

- Problem and most basic algorithm
- Optimizations made to the basic algorithm
- Privacy loss

# Problem and most basic algorithm

**Problem:** Using provable privacy, release the total count and sex by age (23 categories) marginal of population grouped by geography and detailed race and ethnicity.

## **Additional criteria:**

1. We care about keeping relative error low, and statistics with very high relative error should not be published
2. We care about other intermediate breakouts beyond the full sex by age marginal (e.g., sex marginal, sex by age with fewer age categories). If these intermediate statistics are not published directly, we still expect users to derive them.
3. All statistics must be integral, but there are no other consistency requirements.

# Problem and most basic algorithm

**record:**

geo_id	sex	age	race1	race2	...	race8	eth
10000	F	81	1001	5934		Null	1007

Determines mapping to  $\leq 4$   
geographies

Determines mapping to  $\leq 18$   
detailed races/ethnicities

California,  
Egyptian Alone

Total:	0
F, 0-4:	0
F, 5-9:	0
...	...
F, 80-84:	0
F, 85+:	0
M, 0-4:	0
...	...

Orange County,  
Japanese Alone or in Combo

Total:	1
F, 0-4:	0
F, 5-9:	0
...	...
F, 80-84:	1
F, 85+:	0
M, 0-4:	0
...	...

North Carolina,  
Aleut Alone or in Combo

Total:	1
F, 0-4:	0
F, 5-9:	0
...	...
F, 80-84:	1
F, 85+:	0
M, 0-4:	0
...	...

One table per  
population group =

Geography x  
Detailed Race/Ethnicity

# Simple algorithm

- Add discrete Gaussian noise to each statistic, sensitivity is (max geographies per record) x (max race/ethnicity groups per record) x (# statistics each record contributes to) = 144.
- We next present a series of modifications to this basic algorithm to improve error.

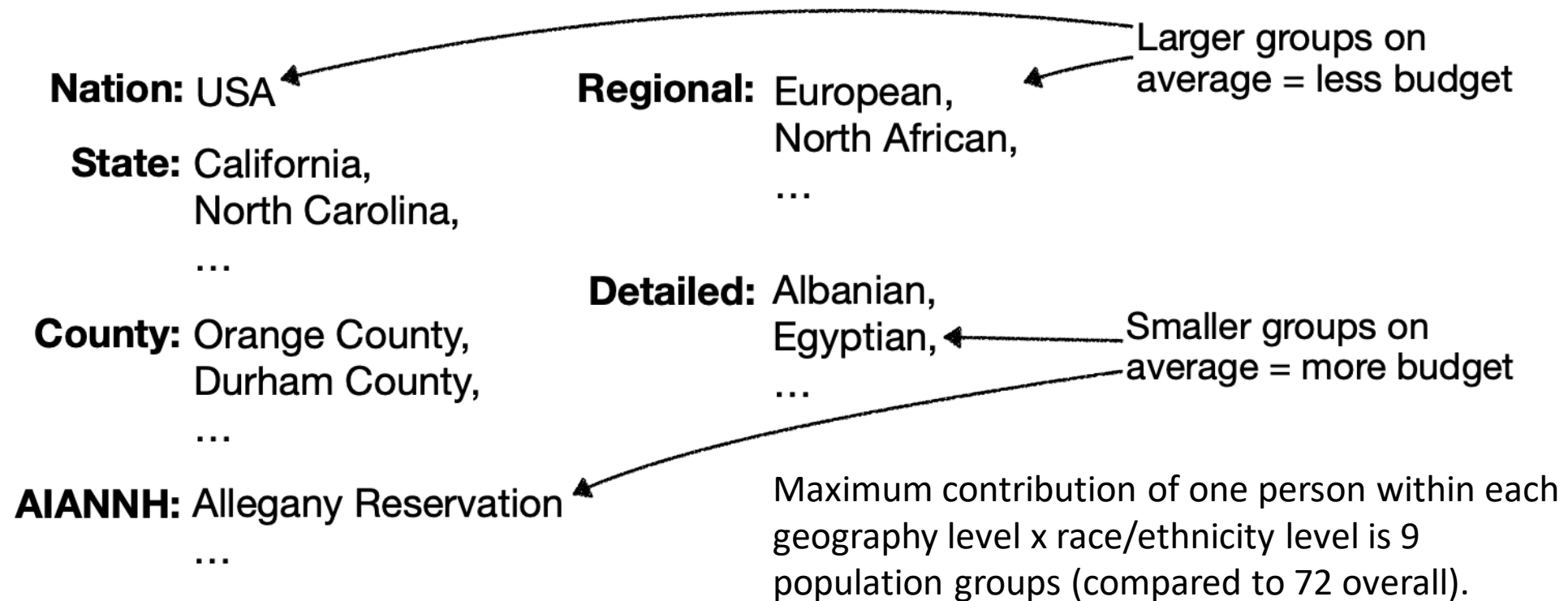


# Differentially Private Algorithms for T1 and T2

- Problem and most basic algorithm
- **Optimizations made to the basic algorithm**
- Privacy loss

# Optimizations made to the basic algorithm

**Optimization 1:** Adjust the privacy budget separately for separate race/ethnicity and geography levels.



# Optimizations made to the basic algorithm

**Optimization 2:** Compute only the most detailed statistics, then aggregate.

- Reduces the sensitivity since each person contributes to at most one count per table.
- Error is higher for aggregated statistics, but these counts are larger.

# Optimizations made to the basic algorithm

## Further Optimizations:

- Choose statistics levels adaptively for some populations groups, based on size of the population group.
- Choose statistics levels up non-adaptively for the remaining population groups, based on 2010 Census data.

# Differentially Private Algorithms for Tables T1 and T2

- Problem and most basic algorithm
- Optimizations made to the basic algorithm
- **Privacy Loss**

# Privacy Loss

- Set target MOE targets for each population group level and compute the scale of the discrete Gaussian noise to hit these targets.
- **Note:** MOE targets for comparison were set based on consultation with the Census Bureau, but do not represent the values that will be used in production.
- Compute the zCDP loss for this noise scale, along with the approximate differential privacy (DP) loss.

# Privacy Loss

Geography level	Iteration level	MOE target	Statistics budget ( $\rho$ )	Total budget ( $\rho$ )
Nation	Detailed	6	0.481	0.534
State	Detailed	6	0.481	0.534
County	Detailed	11	0.143	0.159
AIANNH	Detailed	11	0.143	0.159
Nation	Regional	50	0.007	0.008
State	Regional	50	0.007	0.008
County	Regional	50	0.007	0.008

**Total zCDP privacy loss ( $\rho$ ): 1.41**

**Approximate DP loss ( $\epsilon$  with  $\delta = 10^{-10}$ ): 12.2**

# Update on T3 Household Type and T4 Tenure



# Proposed 2020 Census Detailed DHC – T3 and T4

Detailed race, ethnicity, and American Indian and Alaska Native tribes and villages

- Subjects iterated by detailed race, ethnicity, and American Indian and Alaska Native (AIAN) tribes and villages include:
  - Household type (Table T3)
  - Housing tenure (Table T4)
- We are continuing the development of tables T3 (Household Type) and T4 (Tenure)

# Update on Complex Person-Household Join Tables

# Proposed 2020 Census Detailed DHC - Join

## Complex Person-Household Join Tables

- Subjects requiring a complex person-household join (Join) include:
  - Average household size by age\*
  - Household type for the population in households
  - Household type by relationship for the population under 18 years\*
  - Population in families by age\*
  - Average family size\*
  - Family type and age for own children under 18 years
  - Total population in occupied housing units by tenure\*
  - Average household size of occupied housing units by tenure\*

\*Indicates the table is proposed to be repeated by major Hispanic origin and race groups.

# Development Timeline for Join

## **Beginning in January 2021**

- Introduced to MITRE and Tumult Labs researchers
- Developed PHSafe Tool for evaluating accuracy confidentiality balance

## **Spring – Fall 2021**

- Used PHSafe Tool to propose parameter settings
- Converted from epsilon to rho

## **Winter 2021 - 2022**

- Requested public feedback on the Detailed DHC proposal via the release of the 2020 Census Data Product Planning Crosswalk
- Reviewing feedback from public on use cases for Detailed DHC data
- Developing modeling experiments

## **Next Steps**

- Execute modeling experiments to understand implications of varying parameter settings
- Determine proposed parameter settings
- Continue engaging with public and stakeholders

# DEMO Initial Goals for Join

- Provide same amount of content and geographic granularity as 2010 Census
- Produce accurate data for race and ethnicity iterations A-I that meet use cases
- Produce more accurate and reliable data through modeling

# PHSafe Tool Analysis for Join

- For each table, DEMO can select parameter settings (or options) and the Tool provides the privacy loss budget associated with those settings
- Parameter settings include
  - Geography
  - Iterations A-I
  - Implementation of population thresholds
  - Household truncation threshold
  - Measurement strategy (table granularity)
  - Margins of error

# DEMO Proposed Parameter Settings for Join

- Geography for base tables
  - Nation, state, county, tract, place for all tables
  - AIANNH and Block group for selected tables
- Geography for tables repeated by major Hispanic origin and race groups
  - Nation, state, county, tract
- Measurement strategy
  - Most detailed indentation level
- Continued research to examine:
  - Implementation of population thresholds
  - Household truncation threshold
  - Margins of error

# Public Feedback for Join

- Critical need for Join tables
  - Used by broad population of users for a wide array of purposes
  - Uses that cannot be predicted in advance (e.g., spread of COVID-19)
- Use cases for lower levels of geography
  - Census tract or lower
- Ongoing engagement with stakeholders and partners



# PHSafe: A differential privacy algorithm for Join tables

# PHSafe: A differential privacy algorithm for Join

For each Join table at each population group:

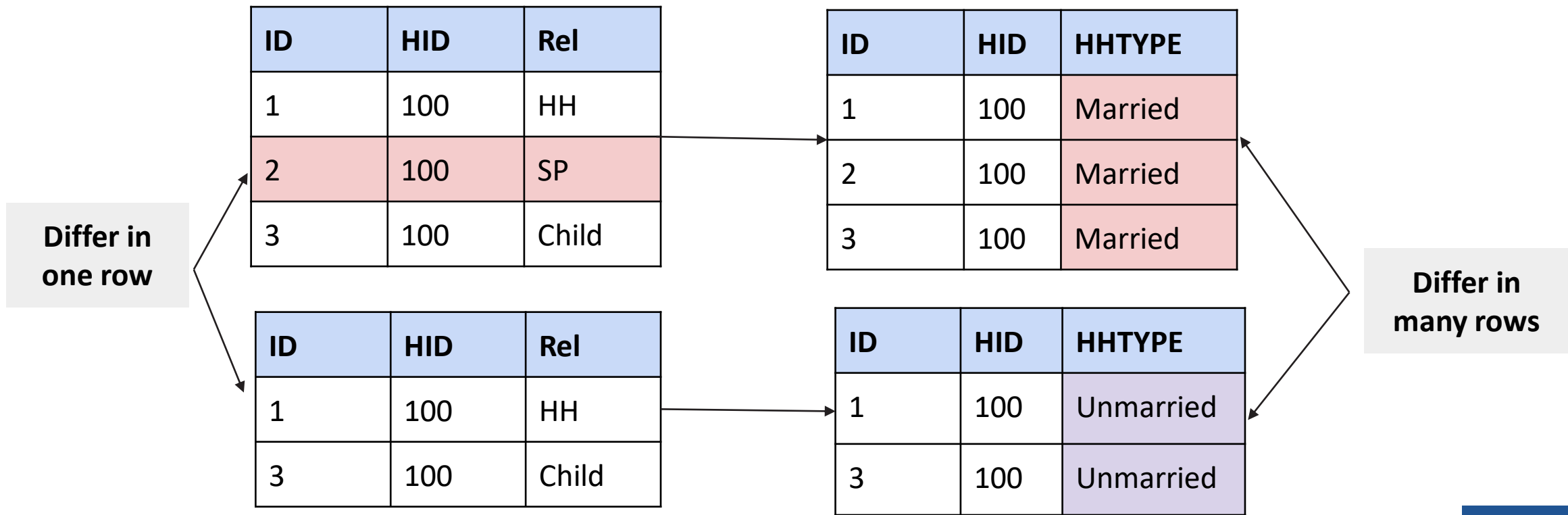
1. **Filter:** restrict to the table universe.
2. **Join with truncation:** Append household attributes from Unit table to each record in Person table, and limit # persons in a household.
3. **Compute noisy estimates:** Add noise to selected table cells.

PHSafe can use Geometric noise (to achieve bounded privacy loss under pure-DP) or discrete Gaussian noise (to achieve bounded privacy loss under zCDP).

# Why limit # persons in a household?

**Sensitivity:** The total change in all query answers due to adding or removing one individual person.

**Higher sensitivity means more error/noise for the same privacy loss.**



# Joins often lead to high privacy loss

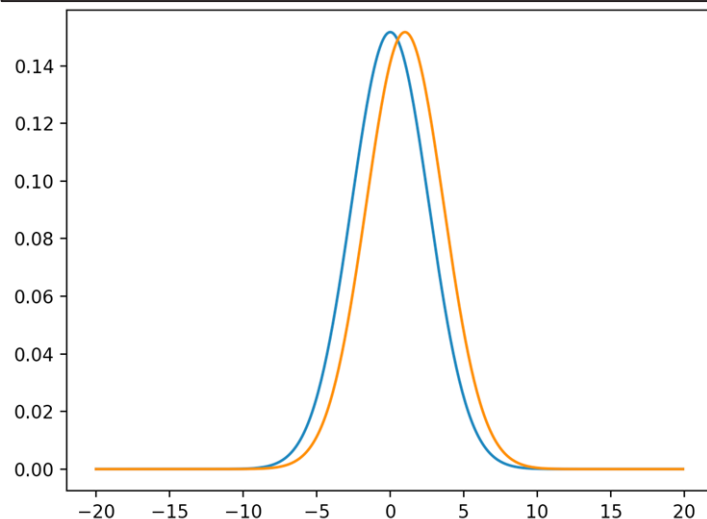
**Sensitivity:** The total change in all query answers due to adding or removing one individual person.

**Higher sensitivity means more privacy loss for the same amount of error/noise.**

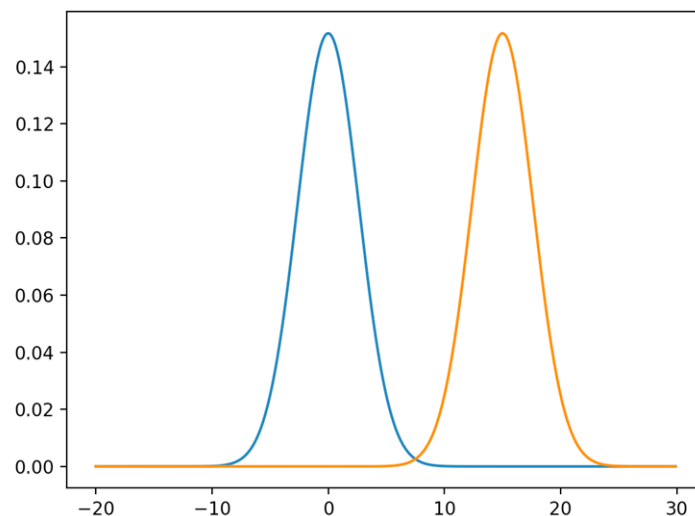
- Join queries are high sensitivity queries. One person can impact the counts in several tables and in magnitude up to about twice the size of the maximum household
- For low MOE targets (e.g., 5 or 10), it is easy for an adversary to infer the sensitive properties of a specific individual.
- Several tables, each of which are iterated for several geographies and race and ethnicity iterations.

# MOE targets & privacy loss for Join

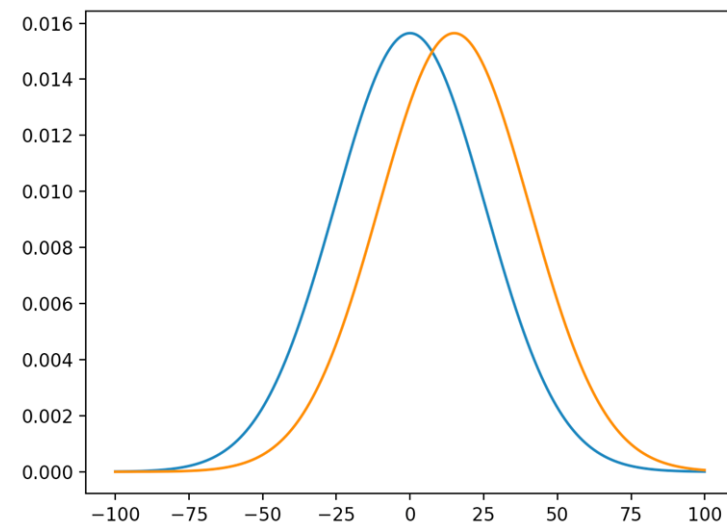
Low sensitivity scenario.  
Max household size = 1  
MOE = 5, low privacy loss



High sensitivity scenario.  
Max household size = 15  
MOE = 5, high privacy loss



High sensitivity scenario.  
Max household size = 15  
MOE = 50,  
moderate privacy loss



# Evaluating privacy-accuracy trade-offs for Join

## 1. Specify Privacy Loss Budget Parameters

Geography	Geography	Enabled	Total Epsilon
	USA	TRUE	2.43
	Region	TRUE	0.54
	Division	FALSE	0.20
	State	TRUE	0.10
	County	FALSE	0.25
	Tract	TRUE	0.10
	Block Group	FALSE	0.20
	Block	TRUE	0.10
	Place	FALSE	0.54
	AIANNH	FALSE	0.10
Total epsilon budget for P30:			3.27

## 3. Set Truncation Thresholds

Truncation threshold:	25
Truncation algorithm:	Random truncation
Drop household above threshold	
Random truncation	

## 4. Choose Measurement

## Error for P30

Region type	Selected	Total epsilon	Level	Cell description	MOE*
USA	YES	0.00	0	Total:	145.20
USA	YES	2.19	1	In married couple household	72.60
USA	YES	0.00	2	Opposite-sex married couple	-
USA	YES	0.00	2	Same-sex married couple	-
USA	YES	2.19	1	In cohabiting couple household	72.60
USA	YES	0.00	2	Opposite-sex cohabiting couple	-
USA	YES	0.00	2	Same-sex cohabiting couple	-
USA	YES	2.19	1	Male householder, no spouse or partner present	72.60
USA	YES	0.00	2	Living alone	-
USA	YES	0.00	2	Living with others	-
USA	YES	2.19	1	Female householder, no spouse or partner present	72.60
USA	YES	0.00	2	Living alone	-
USA	YES	0.00	2	Living with others	-

# Evaluating privacy-accuracy trade-offs (hypothetical example)

		MoEs using 30% CV for 30th P							
		Nation	ST	CTY	TRACT	BLKGRP	PLACE	AIAN	
Household size	8								
P31	Unattributed	200	50	10	5		5	5	
	Race A-G	200	50	10	5				
	Hispanic H-I	200	50	10	5				

Accuracy target (MOE)

Truncation threshold = 8

rho-zCDP Privacy Loss

Table rho = 101

Target rho								
Nation	ST	CTY	TRACT	BLKGRP	PLACE	AIAN	rho/iteration	Table rho
0.01	0.18	4.38	17.54		17.54	17.54	57.18	
0.01	0.18	4.38	17.54				22.11	
0.01	0.18	4.38	17.54				22.11	101.39

High MOE (200) requires low PLB (0.01)

Low MOE (5) requires high PLB (17.54)

# Adaptivity + Modeling -> better accuracy for lower PLB

## Adaptive Algorithm

If (size of pop group is large)

Use the differentially private algorithms to get a detailed direct noisy estimates (so that we can use MOEs of  $\geq 50$ )

(Like in R&E and SafeTab).

## Modeling:

Else // *size of pop group is small*

Use statistical modeling to obtain estimates.  
(Lead by Scott Holan's team at Census)

Minimum MOE	Estimated Rho PLB (P31 - all individuals)
5	~101
10	~35
20	~ 9
35	~ 3
50	1.5 -- 2



# Statistical Modeling for Join

**Scott H. Holan, PhD., Senior Research Fellow, Research and Methodology Directorate**

**Modeling Team: Adam Edwards (MITRE), Joseph Kang (CSRM), Ryan Janicki (CSRM), Kyle Irimata (CSRM), James Livsey (CSRM), Andrew Raim (CSRM), and David Zou (MITRE)**

- Description of the problem
- Phased Modeling Approach

# Description of the Problem

- Develop a principled model-based approach to produce estimates, and associated measures of uncertainty of Join Tables at the county/sub-county level for major race/ethnicities.
- The hierarchical models proposed will make use of an approximating likelihood. Approximations are justified in a working paper (Phase 0).
- Our first modeling priority is to start from the county tabulation level released for each Join Table then to aggregate to higher-level tabulations (Phase I).
- Our second modeling priority for Join Tables is to provide estimates of tabulations at the tract level and for AIANNH Areas (Phase II).
- Model inputs include the differentially private (DP) counts, and their known distributional properties, as well as the uncertainty estimates associated with the DP counts.

# Phase 0 Motivation

- In practice, Phase I modeling presents computational difficulties as the posterior predictive distribution of the underlying "true" values given the unknown parameters and the data is not conjugate.
- To remedy the problem of nonconjugacy, one could estimate the model in Stan, using HMC.
- In principle, this approach provides a viable path forward. However, the nonstandard DP distribution, along with other practical constraints renders this approach infeasible.
- Thus, we approximate the data model under DP using a Gaussian distribution.
- Importantly, for this approach to be viable, we must demonstrate that this approximation results in tabulations that are, on average, more precise than using the DP model directly along with aggregation.
- We do this under three scenarios: iid, regression, and CAR (spatial).
- We demonstrate that our approach provides superior estimates over no modeling.

# Phase I Modeling

Unattributed Table

P31. HOUSEHOLDS BY RELATIONSHIP FOR THE POPULATION UNDER 18 YEARS		
Universe: Population in households under 18 years		
Total:		
	Householder, spouse, unmarried partner, or nonrelative	
	Own child:	
		In married couple family
		In cohabiting couple family
		In male householder, no spouse or partner present family
		In female householder, no spouse or partner present family
	Other relatives:	
		Grandchild
		Other relatives
NOTE: Nonrelatives include any household member not related to the householder by birth, marriage, or adoption except for unmarried partners.		

- Our goal in Phase I is to develop multivariate spatial statistical models that jointly provide estimates of interior cells (i.e., sub-headings under “Own child” and “Other relatives”) in tables like the one above.
- Phase II will consider finer scales of geography (i.e., tract-level).

# Notation - Phase I

- Let  $i = 1, \dots, m$  index the geographic regions.
- Let  $j = 1, \dots, J$  index the sub-headings by Join.
- Let  $Z_{ij}$  be the differentially private (DP) count (*prior to any post processing*) of the number of persons in region  $i$  in a major race/ethnicity group for sub-heading  $j$ .
- Let  $Y_{ij}$  be the true number of persons in region  $i$  in major race/ethnicity group by subheading  $j$ , as determined by the census.

Each DP count,  $Z_{ij}$ , is related to the true count,  $Y_{ij}$ , by the relationship

$$Z_{ij} = Y_{ij} + \xi_{ij},$$

where  $\xi_{ij}$  are independent draws from an appropriately specified differentially private (DP) data distribution.

- Note that in Phase 0, we justify approximating the DP data model using a Gaussian distribution.

# 'Observed' Data/(Possible) Model Inputs

- All DP measurements for all Join Tables at all geographies (State, County, AIAN) as well as the parameters of the noise distributions.
- DP measurements for the Join Tables at the lowest level of tabulation (e.g., county by sub-heading) for each major race/ethnicity group, as well as the parameters of the noise distributions.
- Other publicly available data:
  1. Previous 2010 Census counts
  2. ACS 1-year and 5-year estimates

# Some Goals

- Develop statistical models and computational methods in two parts:
  1. Multivariate spatial county-level model for DP/auxiliary data (Phase I) – e.g., see Bradley et al. (2015, AoAS) and Janicki et al. (2022+, AoAS).
  2. Multivariate spatial models for tract-level and AIANNH areas (Phase II).
- Estimate the true 2020 Census counts  $Y_{ij}$ , more precisely than the noisy DP counts  $Z_{ij}$ . Doing so will allow improvement in privacy/loss budget parameters.
- Provide coarser estimates of other desired tabulations at desired geographies (state, national) through aggregation (Phase I).
- Predict true 2020 Census counts for geographic areas where the DP counts,  $Z_{ij}$ , are not available/too noisy to release (Phase II).

# Questions for CSAC

- What parameter settings (e.g., MOEs, levels of geography) are acceptable to meet data user needs?
- Do you have any feedback on the Tumult Labs methodological work for developing tables T1 and T2?
- Do you have any feedback on the methodology to model Join tables?